

JUL 10 1979

(11/19/5)

HE 19.102: in 2/4 1979

A NATIONAL EVALUATION SURVEY
OF PROJECTS FUNDED UNDER
TITLE IV, PART A OF THE
INDIAN EDUCATION ACT OF 1972

COMPLETED

ORIGINAL

SUPPLEMENT TO
FINAL REPORT

Contract No. 300-76-0397

November, 1978

Prepared for:
U.S. Office of Education
Office of Evaluation and Dissemination
Washington, D.C. 20034
Project Officer: Dr. Emmett L. Fleming, Jr.

15

Prepared by:
Communication Technology Corporation
64 East Main Street
Marlton, N.J. 08053

DESIGN IMPLICATIONS FOR THE IMPACT EVALUATION OF PROJECTS FUNDED BY TITLE IV, PART A, THE INDIAN EDUCATION ACT

This paper was designed to aid the Office of Evaluation and Dissemination (OED) and the Office of Indian Education (OIE) in their deliberations and development of a Request for Proposal (RFP) to conduct an impact evaluation of the Indian Education Act, Part A. The paper was developed in part from findings of the OED sponsored assessment study of projects funded through Part A of the Indian Education Act by Communication Technology Corporation (CTC) and in part from the perceptions and observations of the principal investigator of the study, Jim Fortune. In the paper an attempt was made to separate the recommendations from each of these two sources for the sake of scientific integrity.

Some of the recommendations were taken directly from the results of the three research models which were utilized to test the feasibility of impact evaluation. The Glass model of meta-analysis was used to summarize project achievement test means empirically in an attempt to discuss an overall program effect. A MANOVA model utilizing criteria composed of judgments of project effectiveness taken from multiple sources; namely, the project director, staff and parent committee; was used to test the degree of agreement in judgment across different ratings of project effectiveness. An ANCOVA model using previous year achievement data as a covariant, current year achievement data as a dependent variable and participation in a project with or without objectives in the achievement test area as an independent variable was used to ascertain the sensitivity of tests which were the most frequently used with project instruction.

Other recommendations were taken from the impressions gained by the principal investigator while he worked on all phases of the study including the development of the descriptions of the projects, the conceptualization of the study, the sampling design of the study, the analytic design of the study and the data collection for the study. Rationale for each of these recommendations were reported with the recommendations. All of the recommendations were organized and reported for the six major design categories, namely: conceptualization, sampling, testing, data collection, design, and analysis.

The following discussions of analyses have been extracted from CTC's Final Report on the Study. Therefore, references to tables, as well as the tables themselves, bear table numbers in the original sequence of the Final Report.

The Glass Model

Glass, in his 1976 presidential address to the social conference of the American Educational Research Association, proposed an empirical meta-analysis model for the summary of results of separate research studies. This model required the use of either a pretest-post test control group design or a randomized post test only control group design. The model proposed was metric-free in that it utilized the distribution of normalized difference scores. The achievement test scores collected on an Indian/non-Indian subsample within districts permitted a feasibility test of impact using this model. This model was proposed since different tests were used across districts. This test was run for both reading and mathematics scores for the 4th, 7th, and 12th grades.

For the 4th grade, 39 districts reported usable pretest-post test data (1975-1978 achievement test scores for a number $n = 10$ to 15 Indians and non-Indian students). Table IV-61 illustrates the results of the application of the Glass meta-analysis model to the reading and mathematics achievement tests comparing the distribution of the difference of the "gain" score between Indians and non-Indians. t -coefficients for the mean difference in "gain" scores distributions from zero are $t = -0.75$ and $t = -1.00$ (t having 38 degrees of freedom). It can be noted that there was only a slight non-significant difference in gain, indicating that neither groups' progress was superior to the other. In this case near equal progress denoted a plus in favor of the IEA Title IV Part A projects.

Table IV-62 displays the same comparisons on 7th grade reading and mathematics scores. Again, no significantly different t coefficients, $t = 0.99$ for reading and $t = -1.60$ for mathematics, were obtained with $n = 31$ districts for reading and $n = 28$ districts for mathematics. In this case, the near zero means were of opposite sign, and differential progress could not be assumed.

Table IV-63 displays the same analysis for grade 12, but only four districts reported usable data. The means of the distributions were both positive, but non-significant.

**TABLE IV-61
GLASS META-ANALYSIS
DISTRIBUTION OF GAIN SCORES
GRADE 4**

Variable	N	Mean	Standard Deviation	Corrected SS	T *	PR> T
ZRDIFF*	39	- 0.00001675	0.00013980	0.00000074	- 0.75	0.4590
ZMDIFF*	39	- 0.03628356	0.22641678	1.94805315	- 1.00	0.3233

**TABLE IV-62
GLASS META-ANALYSIS
DISTRIBUTION OF GAIN SCORES
GRADE 7**

Variable	N	Mean	Standard Deviation	Corrected SS	T	PR> T
ZRDIFF*	31	0.00002153	0.00012073	0.00000044	0.99	0.3238
ZMDIFF*	28	- 0.00003978	0.00013146	0.00000047	- 1.60	0.1209

*See note of explanation which follows Table IV-63.

TABLE IV-63
GLASS META-ANALYSIS
DISTRIBUTION OF GAIN SCORES
GRADE 12

Variable	N	Mean	Standard Deviation	Corrected SS	T	PR> T
ZRDIFF*	4	0.00014455	0.00018143	0.00000010	1.59	0.2093
ZMDIFF*	4	0.00009792	0.00007083	0.00000002	2.76	0.0699

*ZRDIFF refers to mean difference between treated group means and control group means in reading.

ZMDIFF refers to mean difference between treated group means and control group means in mathematics.

T represents the results of a standard t-test comparison of a mean with the population mean (μ), $\mu=0$.

MANOVA Test

A second investigation of the feasibility of impact evaluation was made using the general linear models procedures for the multivariate analysis of variance upon multiple judgments of project success. Three MANOVA's were run using a one-way classification and two dependent variable models. The independent variable was whether or not the objective in question was implemented in a given project. The two dependent variables were independent ratings of project success or effectiveness made by staff and parent committee members.

Three MANOVA models were run. They involved the remedial reading objective with ratings of the effectiveness of the projects in overcoming academic problems, the remedial mathematics objective with ratings of the effectiveness of the projects in overcoming academic problems, and the self-concept objective with ratings of the effectiveness of the projects in developing more favorable self-concepts in Native American pupils.

The MANOVA program generated standard analysis of variances tests on each of these dependent variables, and provided the multivariate analysis using Wilks' criterion for the determination of significance. All three tests were reported for each MANOVA.

Table IV-64 displays the statistics for the analysis of the effect of the reading objective. An $F(1,211) \geq 8.89$ for $p < .05$. The parent committee ratings do not approach significance in the univariate analysis. The staff ratings, however, produced an $F(1,211) = 5.89$ which was significant at the $p = .05$ level. The multivariate analysis produced an $F = 3.28$ which was also significant at the $p = .05$ level. An $F(2,210) \geq 3.04$ for $p < .05$. This analysis indicated that projects with reading objectives were rated more effectively in overcoming academic difficulties than were projects which did not have reading objectives.

TABLE IV-64
MANOVA ON READING OBJECTIVE

ANOVA on Parent Committee Ratings

Dependent Variable: PC24

Source	DF	Sum of Squares	Mean Square	F Value
Model	1	0.02003218	0.02003218	0.02
Error	211	206.77809989	0.97999095	
Corrected Total	212	206.79812207		

ANOVA on Staff Ratings

Dependent Variable: Staff 26

Source	DF	Sum of Squares	Mean Square	F Value
Model	1	9.49354652	9.49354652	5.89
Error	211	340.03696992	1.61154962	
Corrected Total	212	349.53051643		

MANOVA

Characteristic Root	Percent	Characteristic Vector PC24	V'EV = 1 Staff 26
0.03128108	100.00	- 0.02371070	0.05631278
0.00000000	0.00	0.06832846	0.00313871

Wilks' Criterion

$L = \text{DET}(E)/\text{DET}(H+E) =$

0.9696674

$\text{EXACT } F = (1-L)/L \cdot (NE+Q-P)/P$

$F(2,210) =$

3.28

$\text{PROB} > F = 0.0394$

An exact analysis as was reported was made using the mathematics objective. No significance was found in any of the three tests. (Table IV-65)

TABLE IV-65
MANOVA ON MATHEMATICS OBJECTIVE

ANOVA on Parent Committee Ratings

Dependent Variable: PC24

Source	DF	Sum of Squares	Mean Square	F Value
Model	1	0.35029598	0.35029598	0.36
Error	211	206.44782609	0.97842572	
Corrected Total	212	206.79812207		

ANOVA on Staff Ratings

Dependent Variable: Staff 26

Source	DF	Sum of Squares	Mean Square	F Value
Model	1	1.33326710	1.33326710	0.81
Error	211	348.19724933	1.65022393	
Corrected Total	212	349.53051643		

MANOVA

Characteristic Root	Percent	Characteristic Vector PC24	V'EV = 1 Staff 26
0.00747117	100.00	- 0.05050694	0.04896887
0.00000000	0.00	0.05178696	0.02654480

Wilks' Criterion

$$L = \text{DET}(E) / \text{DET}(H+E) =$$

0.99258424

$$\text{EXACT } F = (1-L) / L * (NE+Q-P) / P$$

F (2,210) =

0.78

PROB > F = 0.4577

A third MANOVA was run on the self-concept objective. Again, the parent committee ratings failed to produce a significant F ratio; $F(1,211) = 3.78$ whereas $F(1,211) \geq 3.89$ for $p < .05$. The staff ratings, however, produced a significant F ratio; $F(1,211) = 17.36$ whereas $F(1,211) \geq 8.39$ for $p < .05$. The combined multivariate test produced an $F(2,210) = 8.85$ whereas $F(2,210) \geq 3.04$ for $p < .05$. This analysis indicates that projects having the self-concept objective produced more impact in the development of favorable self-concepts in Native American pupils than did the projects which did not have the self-concept objectives. (Table IV-66)

TABLE IV-66
MANOVA ON SELF-CONCEPT OBJECTIVE

ANOVA on Parent Committee Ratings

Dependent Variable: PC27

Source	DF	Sum of Squares	Mean Square	F Value
Model	1	3.08789274	3.08789274	3.78
Error	211	172.18910256	0.81606210	
Corrected Total	212	175.27699531		

ANOVA on Staff Ratings

Dependent Variable: Staff 29

Source	DF	Sum of Squares	Mean Square	F Value
Model	1	24.82755884	24.82755884	17.36
Error	211	301.80154915	1.43033910	
Corrected Total	212	326.62910798		

MANOVA

Characteristic Root	Percent	Characteristic Vector PC27	V'EV = 1 Staff 29
0.08431905	100.00	0.01254268	0.05385344
0.00000000	0.00	0.07936695	-0.02799006

Wilks' Criterion

$$L = \text{DET}(E)/\text{DET}(H+E) =$$

0.9223778

$$\text{EXACT } F = (1-L)/L \cdot (NE+Q-P)/P$$

F (2,210) =

8.85

PROB > F = 0.0002

The ANCOVA Model

As a third alternative to a feasibility test, four analyses of covariance were run on the achievement tests with the greatest frequency of use. Six were planned but there were not enough reading and mathematics scores for the 12th grade to conduct an analysis of covariance. Hence, tests on reading and mathematics achievement scores in the 4th and 7th grade were conducted using SRA's grade equivalents, the most frequently reported scores.

Tables IV-67 through IV-70 report the results of these two-way analysis of covariance tests. SRA grade equivalent scores 1976 were used as the dependent variable, 1975 SRA's grade equivalent scores were used as a covariable, the subjects were grouped as to whether or not they were in a project having an objective appropriate to the test — reading to reading scores, mathematics to mathematics scores, and were also grouped as to Indian/non-Indian. In all four tests — two for the 4th grade and two for the 7th grade, the covariable explained most of the variance, hence no differences between groups or any effects were found. In all cases means were similar enough to post no differences.

TABLE IV-67
ANCOVA — GRADE 4
SRA TEST — READING

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Covariates	378.123	1	378.123	457.997	0.0
R11	378.123	1	378.123	457.997	0.0
Main Effects	1.407	2	0.703	0.851	0.428
OBJ1	0.245	1	0.245	0.296	0.587
RACE	1.166	1	1.166	1.413	0.236
2-Way Interactions	0.444	1	0.444	0.538	0.464
OBJ * RACE	0.444	1	0.444	0.538	0.464
Explained	379.974	4	94.993	115.060	0.000
Residual	221.261	268	0.826		
Total	601.234	272	2.210		

TABLE IV-68
ANCOVA – GRADE 4
SRA TEST – MATH

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Covariates	266.259	1	266.259	341.576	0.0
M11	266.259	1	266.259	341.576	0.0
Main Effects	0.777	2	0.389	0.498	0.608
OBJ1	0.567	1	0.567	0.727	0.395
RACE	0.215	1	0.215	0.276	0.600
2-Way Interactions	0.000	1	0.000	0.000	1.000
OBJ1 RACE	0.000	1	0.000	0.000	0.38
Explained	267.036	4	66.759	85.643	0.000
Residual	207.347	266	0.780		
Total	474.383	270	1.757		

TABLE IV-69
ANCOVA – GRADE 7
SRA TEST – READING

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Covariates	142.728	1	142.728	29.531	0.000
R11	142.728	1	142.728	29.531	0.000
Main Effects	1.227	2	0.614	0.127	0.881
OBJ1	0.002	1	0.002	0.254	0.615
RACE	1.226	1	1.226	0.254	0.615
2-Way Interactions	11.534	1	11.534	2.386	0.124
OBJ1 RACE	11.534	1	11.534	2.386	0.124
Explained	155.489	4	38.872	8.043	0.000
Residual	1304.972	270	4.833		
Total	1460.461	274	5.330		

TABLE IV-70
ANCOVA — GRADE 7
SRA TEST — MATH

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Covariates	134.947	1	134.947	36.755	0.000
M11	134.947	1	134.947	36.755	0.000
Main Effects	1.739	2	0.870	0.237	0.789
OBJ1	1.316	1	1.316	0.358	0.550
RACE	0.352	1	0.352	0.096	0.757
2-Way Interactions	6.841	1	6.841	1.863	0.174
OBJ1 RACE	6.841	1	6.841	1.863	0.174
Explained	143.527	4	35.882	9.773	0.000
Residual	906.873	247	3.672		
Total	1050.440	251	4.185		

Summary of Results

Given the slight potentiality of results in reading in all three tests, given the weakness in number of districts for which usable data were available, the apparent pretest differences between control and treatment groups, and the absence of a common metric across districts, the feasibility of impact evaluation in reading appears favorable. A standard achievement test like the Metropolitan, enhanced by equivalent comparison groups, controlled testing and nearly complete data returns could be found to discriminate project effects in reading. Findings in mathematics, however, do not suggest that impact evaluation in mathematics is feasible. These negative findings may be due to the fewer number of mathematics projects and the smaller number of participants in addition to the large amount of missing data, the absence of a common metric and the large differences between comparison groups.

MANOVA results supported by a small sample of local evaluation results in favor of self-concept improvement suggest a potential impact evaluation in the area of self-concept development could be made if a suitable instrument can be found. The use of multiple-party ratings of project effectiveness are suggested as potential validation evidence in the design of any impact evaluation request for proposal.

Recommendations in Regard to Conceptualization

Due to the absence of more than half of the achievement test data across districts and due to the existence of only 20% of the districts possessing data of adequate quality for analysis, the test of feasibility of impact is inconclusive.

Results of the Glass model analysis suggested that if more equivalent comparison groups (Indians in the project and Indians not in the project) were obtained and a controlled testing situation were imposed, that impact evaluation may be feasible in reading. The Glass meta-analysis model was run over all projects having achievement data available. Equal numbers of non-Indian students of matching grade level were selected randomly as a control group. In many instances the pretest achievement scores (from the previous year) were larger for the non-Indian control group than for the Indian project group. Again also was the case that all projects did nothing to try to improve reading and in some cases improvement of reading did not rate a high priority in regard to project goals. The results of the Glass model analysis did not indicate significant ($p < .05$) positive results in that Indian children did not progress at a more rapid rate than did the non-Indian children progress. In one sense, this finding is showing program or project effects in that Indian children from previous scores should be progressive at less of a rate than is observed for non-Indian children. Hence, the near years difference in rates of progress is a positive effect. The MANOVA model produce significance ($p < .05$) in reading programs in regard to project ratings. Although this evidence of impact is not based upon achievement data, the results tend to lend credibility to the conclusions made concerning the Glass model results.

Although no differences were obtained in the ANCOVA tests, the groups were extremely small and differences between pretests of comparison and treatment groups were large. There were observable tendencies of increased gains in favor of Indians as opposed to non-Indians in reading but not in mathematics.

These results tend to suggest two recommendations in regard to Impact Evaluation Conceptualization:

1. The impact evaluation model must identify a means of obtaining near-equivalent Indian control groups. This may be accomplished through analysis of non-participants within the same funded school district and at the tested grade level.
2. The impact evaluation must be conceptualized to address the most frequently obtained objectives and must seek at all times to overcome the problem of small groups of treated or served pupils at any given grade level or within any given project objective. Reading was one of the most frequently entertained objectives and was usually sought with other objectives such as affective objectives.

Five additional factors should be considered in the conceptualization of the impact evaluation. These include:

1. Generally, the Indian Education Program is a culturally based and an affective-domain based program in which projects seeking only cognitive objectives were rated more frequently as unsuccessful than were projects seeking objectives across all three domains. The provision of medical, psychological, and counseling services appeared to supplement cognitive objectives but not to the degree these Indian studies and self-concept objectives did. Conceptualization of the evaluation must seek to control for the numbers of supporting objectives are entertained.
2. Many of the school districts with large numbers of eligible students are either small or intermediate sized districts in regard to total enrollment and for the most part they fall in the range of having average sized per pupil expenditures. The conceptualization of the impact evaluation should include considerations of this phenomenon by describing the capability of each participating district and by reducing data-provision burdens on the district's staffs as much as possible.
3. Indian people believe that there are some unique, culture based differences in the needs being addressed by the projects. Although there was no evidence to suggest that improper needs were being addressed, their belief in the uniqueness of Indian needs has led to an Indian community mistrust of non-Indian researchers, non-Indian based tests, non-Indian data collectors, written questionnaires and standard research and evaluation methodology. This phenomenon of mistrust must be considered in the conceptualization of the study through the use of Indian input in study design, careful and simple explanations of study objectives and strategies, and patience in the design of data collection systems.
4. The absence of a concentration factor and the level of project funding which provides a per eligible student entitlement of \$180 financial and some \$120 below the suggested level of impact in ESEA Title I evaluation studies, interacting with the wide range of needs and target group disparities suggested that impact may vary greatly across projects, thusly requiring a highly controlled study.
5. Although the Indian Education Projects may address unique and supplemental goals within a project, the same goals may be sought for other Indian students in other districts through other federal programs. The conceptualization of the study must address need and receipt of a given subject and then ascertain funding source and level. Singular program comparisons will not account for competing hypotheses generated by alternative funding sources.

Recommendations in Regard to Sampling

Results of the assessment of project effectiveness in regard to differences of projects and of project effectiveness across the stratification variables in the sample, suggest that two of the three variables used in the sample frame were associated with project effectiveness. Association with project effectiveness was found for the composite characteristic variable, urban-rural high and low density environments and for the trichotomous variable created on size of grant. Geographic location did not appear to be associated with effectiveness ratings although there were some differences in project objectives across the cultural regions.

Experience gained through the deliberation over appropriate and efficient sampling designs for the CTC study led to the following considerations in the choice of a sampling design for impact evaluation. The considerations and recommendations include:

1. The stratified cluster sample with unequal fractions selected proportional to differences in average strata size does not provide the efficiency in sampling design for the Indian Education Program as it does for other familiar programs such as ESEA Title I. In the case of ESEA Title I, the stratification variable is based upon total enrollment where the range of cluster sizes is very large and where 80% of the students are contained in the two largest of the five strata conventionally used. In the case of Indian Education the stratification variable should be based upon Indian enrollment which possesses the same relationship to Indian Education funding as enrollment does to ESEA Title I funding. However, in this case the range of cluster sizes is not as large as is found when using total enrollment and only 75% of the students are contained in the three largest of the five strata investigated. Hence either a larger sample must be studied or a new basis of sampling efficiency be found. This might be provided by the number of objectives sought by each project.
2. The degree to which anticipated small sizes of impact can be expected to be found mandates use of large subsamples (thirty as a minimum) and the selection of robust analytic models. Weighted samples provide difficulty in interpreting analytic models based upon the least-squares principle. To maximize the potential of gaining clear-cut and meaningful internal validity, the designer of the impact study should make careful consideration of self-weighting samples.

Recommendations in Regard to Testing

1. The Metropolitan Achievement Test in reading was shown to discriminate across projects with reading objectives, but not across projects with objectives in mathematics. Likewise, more significant trends across program participants were found in reading, but not in mathematics. If appear more

complete data could be gathered, then a standardized test in reading could be found to discern program effects. Similar hope is not maintained, however, in the area of mathematics, where there are fewer projects and project participants.

Although the finding of significance in reading programs was made on project ratings, this result offers some credence to the Glass results. Here, the completeness of the data is much larger so as to include more than 90% of the data. Again, no similar hope for mathematics is maintained.

2. Given that a suitable self-concept inventory could be identified (only five districts reporting using self-concept scales and three of these showed positive gains), the MANOVA on ratings suggests the potential of impact evaluations in the self-concept area.

In the design of the study several testing requirements became clear.

1. The wide range of type of tests used, methods used in recording and scoring the tests, incompleteness of archival data files, and absence of supporting testing information such as irregularities and time of testing make the dependence upon archival test files unwise. The designers of the impact evaluation must plan to control the testing so that: (1) they can assure common metrics across projects, (2) they are not faced with small samples of test takers for any given test, (3) they do not have to address the off-level testing problems suspicioned in the other study, (4) they do not have to contend with missing data and absence of descriptions of test-taking conditions and time and (5) they have control over the analytic unit selected for analysis. The NCE unit concept proposed in the new ESEA Title I evaluation system may be considered as an appropriate metric.
2. Control of the reaction of the Indian community and improvement of the relevance of the study design to the program can be gained through the expansion of the definition of impact to focus upon primary program emphasis. The designers of the impact study should consider measures of self-concept, attitude toward school, and social competence as other means of discerning educational impact. Ratings of project effectiveness could be used as back-up information.

Recommendations in Regard to Data Collection

1. The designer of the impact evaluation should devote high consideration to the strategy of collecting saturation data about a given objective for which impact is being sought. OMB reduced other federal program information from the prior studies data file prohibiting the completion of a total saturation analysis of any specific project objective. The absence of the availability of the results of the saturation analysis in areas such as mathematics prevented any attempted probes to explain the absence of internal validity.

2. Data collection should be made through site visits to maximize the response rates and the quality of the data especially in regard to gaining Indian responses. Although most districts were extremely cordial and cooperative, due to the across culture nature of the information mixed (Indian, non-Indian) data collection teams appear to produce the most trouble free method of data collection. The lead man concept appeared to contribute to the efficiency of the data collection and the computer aided receipt control system and the WATS trouble shooting system both appeared effective.
3. Multiple responses across staff members and parent committee respondents were for the most part very similar. A single random respondent from each group appears to provide adequate representation.

Recommendations in Regard to Study Design and Analysis

1. Pretest and post test control group design should be considered so as to control for the time between testing and to create a fair program comparison. Creativity in the choice of a control group will be essential. It is recommended that the control group should also be Indians.
2. Several methods of independent variable construction using the interface of setting variables, project variables, and target group characteristic variables should be considered as a means of creating variables to describe treatments.
3. Data analysis models should include the capability to block across participation/non-participation in the project objective of the impact area, across different levels of multiple objective projects and across different levels of Indian density.
4. Project impact evaluation should focus upon a selected sample of three or four grade levels and upon a sample of three or four impact areas discerned from the frequency of project objectives.
5. Program analysis within projects should focus upon individual impact analyses of individual projects and discriminant analyses of descriptive characteristics between successful and unsuccessful projects as defined by the presence or absence of impact internal validity.
6. Program impact should be investigated using project level units of analyses so as to avoid ecological inference which is indicated by high within project variance.
7. Regression-discontinuity analytic models and other regression-based analytic models should be considered as the basic analytic model for program level impact analysis and for individual project level impact analyses.